Identifying Medications that Patients Stopped Taking in Online Health Forums

Jason H.D. Cho[*‡§], Tony Gao[†‡] and Roxana Girju[*†‡]
[*]Department of Computer Science
[†]Department of Linguistics
[‡]University of Illinois at Urbana-Champaign
Urbana, Illinois 60801
[§]@WalmartLabs
Sunnyvale, California, 94086
{hcho33,tonygao2,girju}@illinois.edu

***Abstract*—Patients may stop taking medications after a certain point for various reasons, such as severe side effects, prohibitive costs, or ineffective treatments. Being able to analyze the reason patients stop taking medications is very important to medical practitioners, for example, who can come up with new treatment plans, prescribe different medication if there are side effects. In this paper, we focus on online health forums and define the problem as a binary classification task (i.e., if a patient has stopped taking a medication or not). We chose to focus on health forums here since these are the platforms usually patients go to ask for support online. We propose linguistics features of various complexity and present an in-depth analysis of the results which give us new insights into the task at hand.**

I. Introduction

Medication non-adherence is a huge issue in the health community where up to about half the patients do not take medication as prescribed [1], [2].Not surprisingly, non-adherence is responsible for an estimated burden of 337 billions dollars per year in direct and indirect health care costs [3]. Analyzing why patients do not adhere to prescribed medications has a huge impact on cutting health care costs in the long run.

In analyzing medical non-adherence, the first step is in understanding why people have stopped taking a given medication, a subset of non-adherence. According to a previous study [4], there are four reasons why patients stop taking medication:

1. Experience of loss by the use of medication (adverse drug reaction)
2. Personal *meanings* associated with taking medication (taking anti-depressants means admitting one is depressed)
3. Different *feelings* evoked by the process of taking medication (emotions, such as disgust or humiliation at the thought of taking the medication)
4. Perceived changes in payoff matrix (efficacy of medication)

However, there are many limitations to this study. For example, the experiments were conducted in a controlled session where researchers asked participants in-depth questions. Due to the formal environment/setting in which the study was conducted, patients may not have expressed themselves as freely as they would have in an informal setting. Moreover, it is very difficult to scale up the research and generalize it to different demographics in such a formal environment.

A possible solution to this problem is to make use of online social media. This has been an active medium for various healthcare tasks such as epidemiologic studies which looked at drug adverse reactions [5], [6], [7]. Social media is also appealing due to its scale - i.e., it is possible to conduct studies on a large number of probems with imput from many people/patients around the world. Furthermore, most of the research seems to indicate that the findings in this medium are consistent with the real world findings as well.

In this work, we tackle the issue of medication use. In particular, we want to analyze why people have stopped taking medication and, as a first step we focus on identifying whether a person has stopped taking a given medication or not - a binary classification problem given a medication of interest. We note that this is not the only approach on how a medication may be used - for example, patients may indicate that they are currently taking the medication, or may simply be asking questions about a given medication. We consider these intents as 'Others' category and focus primarily on identifying whether a person has stopped taking a given medication or not. We used user messages in health forums as appropriate data and genre because people tend to give more descriptive reasons on why they may have stopped taking medication compared to micro-blogs despite potentials for bigger amounts of data. We have identified feature sets that were successful in similar tasks, and adapted these to suit our problem at hand. We then analyzed these features to see how different complex features perform with regard to the identification process.

Our contribution is as follows:

1. To the best of our knowledge, we are the first to tackle the problem of whether a person stopped taking medication or not in the online health domain.
2. We adapted existing features that were successful in previous health status tasks (we denote these as informative baselines features), and proposed new complex features for the problem domain. We further analyzed how much each of the different feature types contribute

to the classification performance.

3. Our experiments also provide interesting insights not only to the language technology community, but also to health practitioners and the healthcare insurance industry. For example, when patients stop taking anti-depressants, they generally seems to have tried more than one medications, experience that generates emotions such as disgust and sadness.
4. We make our new, reliably annotated dataset publicly available. (website will be provided in the conference proceedings).

II. Related Works

Previous research has focused mainly on identifying the reasons for which people have stopped a health related behavior. A famous example is an I2B2 challenge where the task was to identify patients' smoking status from medical discharge records [8]. The competition results show that a combination of good keywords, lexical and semantic features is important in improving the classification accuracy [9]. Although related, our work differs in two respects. First, we focus on health forums where sentence structures tend to be less grammatical. Second, there may be more than one type of medications that are mentioned in a given text. This makes the task of identifying which medication the patient had stopped taking, and which medication the person is currently taking a lot more challenging.

Another work, perhaps most similar to ours, analyzed health forums about drug addictions [10]. They proposed a conditional random fields model to show how patients go from being addicted to a drug, to weaning out, to withdrawal or recovery. The biggest difference between this work and our work is that we focus on prescription drugs whereas they focus specifically on drug addiction and recovery, so the problem setting is a bit different. Furthermore, they focus more on the individual's addiction phases, whereas we are more interested in the status of medication, i.e., the focus is more on the medication than the patient.

Aggregating health information from social media is another popular area of research. One line of work focuses on finding adverse effects of a given medication [11], [7], [12], with the goal of identifying undiscovered side effects from health forums. They assume that side effects can be modeled by topic modeling in this line of research. Similarly, summarizing drug usage from social media is a growing line of research where some works model how different demographics use drugs differently [13], [14] while others focus more on pharmacovigilance [15], [16].

III. Dataset and Experimentation Settings

In conducting our experiments, we first collected a dataset in which we labeled whether a patient had stopped taking a given medication or not. Unfortunately, we couldn't find any (available) dataset annotated with this kind of information, so we annotated the dataset ourselves. We first crawled a

Drug	Zoloft	Paxil	Cymbalta
Agreement	0.81	0.75	0.85

TABLE I: Cohen's Kappa coefficient for Inter-rater agreement for the three different medications considered

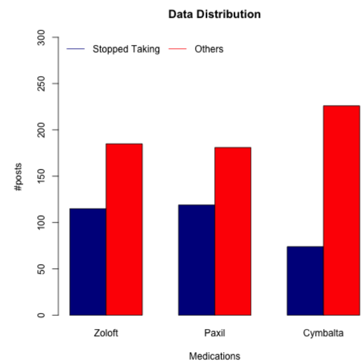

Fig. 1: Number of labels for different medications

depression forum[1]. The rational is that depression is one of the most prevalent conditions that affects a wide range of demographics [17]. By manual inspection of a few forum messages, we decided to focus on three popular depression drugs (Zoloft, Paxil and Cymbalta) as our target medication to annotate whether the patient had stopped taking the medication or not. For each medication, we chose an opening post which contained the medication of interest in a given thread and determined whether the person had stopped taking the medication or not. Notice that our task is to classify, for the 'Stop' label, whether the person has stopped taking the medication of interest. The Others label refers to all the other situations, such as they are currently taking the medication, or wanting to learn more about the medication, or citing previous research conducted on the medication. Two annotators, one student in Linguistics and one in Natural Language Processing, labeled 300 forum posts for each medication, for a total of 900 labeled forum posts. These forum posts were randomly selected from the crawled dataset. Next, we calculated the agreements using Cohen's Kappa coefficient, shown here in Table I. The data distribution is shown in Figure 1.

For any forum posts that the labelers disagreed on, there was an adjudication process to determine the final labels. Change of dosage was often a source of confusion, for example *I was on Zoloft 100mg and then double and then i even eat a crocus plant...*. Confusions were especially evident when there were multiple mentions of changes of dosage. These were adjudicated after re-reading the forum posts and then agreeing upon the final label. A subset of these errors occurred when patients were weaning off medications. While the intents of these were often to stop taking the medication, unless it is clear from the text that they ultimately stopped taking the medication, these were generally labeled as 'Others'. Another source of confusion is when the forum post was lengthy as

[1] www.healthboard.com

there was a lot of information the labelers had to absorb (they had to read the text multiple times).

For all of the experiments, we used Support Vector Machines[2] with 10-fold cross validation. We chose SVM because it is generally considered to be a good classifier on sparse datasets.

IV. Features

We experiment first with informative baseline features. For this, we identified features from the research literature on projects similar to ours. We then came up with additional features to address common errors the informative baseline features made. Unless otherwise annotated, we considered all our features equally likely, thus having a weight value of one.

A. Informative Baseline Features: Definitions and Examples

Keyword Pivoted Word Features: In the I2B2 challenge, using a set of good keywords was one of the most popular and effective approach [9], [18]. One of the authors proposed extracting sentences which contain a set of keywords, and ran the K-Nearest Neighbor model on these sentences [18]. Another group proposed using windows of text up to 100 characters instead [19]. Following the spirit of these works, we experimented by curating a list of keywords: *stop, quit, drop, try, wean, switch*. We select a sentence if it has one of the curated keywords and the medication we are interested in, and then extract all the words as our feature set, using TF-IDF to assign weights to words.

Odds Ratio Features: MacLean, et. al., 2015 [10] proposed using odd log ratio for extracting keywords on **post** level as opposed to **sentence** level. This was feasible because the authors were interested in inferring the status of one medication in a given forum post, as opposed to potentially inferring multiple instances of the medication which is what we focus on. Thus, we adapt the odds ratio in such that we only take sentences that either have the medication of interest, or are those that precede, or succeed the drug. The ratio is a measure of strength of association and is given by:

$$OR(t,s) = \frac{f_s(t) \cdot f_{\overline{s}}(\overline{t})}{f_s(\overline{t}) \cdot f_{\overline{s}}(t)} \quad (1)$$

where s is the state (such as Stop taking or Others), t is the term we are interested in calculating odds ratio, and $f_s(t)$ is the number of sentences with state s that contain t. $f_s(\overline{t})$ is the number of sentences that does not contain the term t for state s, and $f_{\overline{s}}(t)$ is for any other states that are not s, how many times the term t appears. Similar to a previous research work [10], we also retain odds ratios which are greater than 2, and set the ratio as the feature weight for the odds ratio word. Furthermore, we do not include words that appeared less than 10 times in our corpus. We compare the results of calculating the odds ratio on a post level versus sentence level in our experiments, and indeed, we see that sentence level odds ratio improves over that of post level.

Curated N-gram Pattern Features: Using carefully curated rules was shown to be effective in detecting whether patients stopped smoking or not in medical notes [20]. This kind of work in particular proposed a set of seven rules to indicate whether a person may be a smoker or not. We adapt this approach and propose an automated rule generation framework which does not involve the stop keywords. In particular, we noticed that when a patient starts a sentence with a verb in past tense, followed by a preposition and a medication, it is likely to be an indication that the patient is not currently taking the medication. This may or may not be followed by a duration/temporal keyword. As an example, in the sentence *I was on Zoloft for 3 months*, we see that the past tense verb (was) is followed by a preposition and a medication name. Similarly, when a sentence starts with a present tense verb, it may be an indication that the patient is currently taking the medication, for example, *I am currently taking Zoloft* (adverbs are important here as well). We further note that patients may mention that they have switched from medication A to medication B. We add these as a feature set as well. More formally, we encode these rules as

[VBD/VBP] [PREP] [MED]

[VBD/VBN] [FROM] [MED] [TO] [MED]

where VBD is past tense verb, VBP is a present tense verb and VBN is verb participle. PREP and MED correspond to preposition and medication, respectively. We also replaced pronouns with the medication name if only one medication was mentioned in the previous sentence (i.e., we resolved the coreference). We further allowed the pattern words to appear few words away from each other, i.e., a PREP may appear 3 words after a VBD or a VBP.

Time Features: Mentions of time expressions has also been a very popular approach in previous research, either as part of a carefully orchestrated set of rules [20] or by utilizing a system that can distinguish expressions that indicate current time, recent past, and distant past [21]. One previous work [20], for example, used time tokens as part of the carefully orchestrated rules to improve the performance of the health status detection. Motivated by these lines of work, we used the rule generation framework from the previous section, and added temporal words at the end. This framework would capture the sentence *I was on Zoloft for 3 months* as having a temporal word since the word months appeared after the rule *[VBD/VBP] [PREP] [MED]*. Examples of temporal words are: *currently, now, day, month, week, year*.

Emotion Features: Finally, we experimented with emotion features as well. One of the reasons for which people stop taking medication is the feeling that may be evoked by the act of taking the medication [4]. For instance, if they do not like the idea of taking the medication itself (hatred), or the medication reminds them of some other events which make them feel sad, patients may stop taking the medication. Thus, we used SenticNet [22] which makes use of the idea of hourglass of emotions in modeling the emotions of interest here.

[2]We used Scikit-learn for our experiments. We ran validation on the cost parameter C to set the correct parameter, and set the parameter dual to false so that we can analyze the learned weights that we analyzed later in this section. We did not notice any significant differences in the performance results.

Informative Baseline Features		
Feature type	**Example sentence**	**Example features**
Keyword pivoted features	[S]o i was flusterated, and just **stopped** taking my **paxil** for a few days.	TF-IDF weighted words that appears in this sentence
Odds ratio features	I quit the Cymbalta cold turkey and am now going through withdrawal.	quit, cymbalta, cold, turkey
Curated n-gram pattern features	I've been off the paxil for a while.	've/VBP off/IN MEDICATION
Time features	I was on Zoloft about 3 years ago	was/VBD on/IN MEDICATION TIME_WORD
Emotion features	Effexor gave me hives, Zoloft gave me horrible nightmares, and Wellbutrin gave me horrible hives after 2 1/2 weeks on it.	sensitivity: -0.51, attention: 0.21, aptitude: -0.52, pleasantness: -0.28

Complex Features		
Feature type	**Example sentence**	**Example features**
Dependency relations features	I stopped taking the Cymbalta immediately and have now been off it 7 days.	... , (I, PRP) nsubj (stopped, VBD), (taking, VBG) xcomp (stopped, VBD), (Cymbalta, NNP) dobj (taking, VBG), (the, DT) det (Cymbalta, NNP), ...
Frequent sequential pattern features	He put me on effexor.	put on MEDICATION_TOK, me on MEDICATION_TOK
Co-reference resolution features	I've been clinically depressed for the past few months and Effexor hasn't helped me much. True, I've been taking only 75 mg of the **medication**...	Change 'medication' to 'Effexor' and apply all the other features

TABLE II: Example Features used in our experiment

Dimension	Sensitivity	Attention	Pleasantness	Aptitude
Positive	Anger	Anticipation	Happiness	Trust
Negative	Fear	Surprise	Sadness	Disgust

TABLE III: Emotions used as features

Method	**Precision**	**Recall**	**F-1**	**Accuracy**
Passage OR	0.540	0.523	0.528	0.680
Sentence OR	0.641	0.562	0.593	0.737

TABLE IV: Performance of odds ratios

The work organizes emotions around four independent, yet concomitant dimensions (sensitivity, attention, pleasantness, aptitude), whose different levels of activation make up the total emotional state of mind. Each of the dimensions may have positive or negative polarity which maps to different set of emotions. As an example, a positive pleasantness may indicate happiness, whereas a negative one may indicate sadness. Table III shows all 8 emotions we used from SenticNet.. If a sentence contained the medication of interest, we used this emotion ontology and then normalized by the log of the number of words in the sentence. This gave us polarity scores for each of the four dimensions. All of the features described in this section are summarized in Table II.

B. Informative Baseline Features: Experiments and Analysis

We were first curious about whether it would be beneficial to use words taken from the sentence level or from the passage level. Sentence level odds ratios were calculated using words taken only from those that contain the medication of interest, whereas post level ratios made use of the entire post. Our results are shown in Table IV.

We see that sentence level odds ratio outperforms the passage level one. This makes intuitive sense - in a given post, patients are likely to talk about myriads of topics and limiting the odds ratio calculation to the sentence which contains the medication of interest allows us to utilize only the words that are relevant to the medication itself. We show the top 10 highest ranked words for both sentence and passage odds ratios in Table V. In the case of sentence odds ratio, many of the words make intuitive sense. For example, 'cold turkey' is an often used expression when a patient stops taking medication suddenly. We see both 'cold' and 'turkey' in the top 10 highest ranked odds ratios. On the other hand, passage odds ratios do not quite indicate whether a person has stopped taking medication or not. It does, however, show that someone had something to do with medication ('baby', 'dad', 'mom', 'man' and 'depressant'). This can be because in generating our dataset, we chose forum posts which contain the medication of interest.

Sentence OR	stopped, attacks, cold, point, upset, lbs, thinking, saw, made, turkey
Passage OR	baby, become, dad, yrs, mom, failed, per, helping, met, depressants, man

TABLE V: Top 10 highest ranked odds ratio words

Next, we were curious to see how much each additional different feature contributed to the performance of the task. Based on the previous results from experimenting with different types of odds ratio, we opted for sentence level odds ratio as the baseline method. Our results are shown in Table VI where we experiment by taking one feature out from all the feature sets that we have described in the previous section.

Other than the odds ratio features which we used as the baseline, curated n-gram pattern features seemed to have the biggest impact on the performance. This indicates that curated n-gram pattern features correctly captured many of the relevant phrases. Keywords played a big factor as well. Along with odds ratio, keyword based TF-IDF features captured the context in which the medication word appeared. Emotions also had some impact in the classification task. This seems to reiterate one of the hypotheses in our introduction where patients stop taking medication because of the feeling that

Method	Precision	Recall	F-1	Accuracy
Odds Ratio (Baseline)	0.641	0.562	0.593	0.737
Informative Baseline	**0.673**	**0.640**	**0.650**	**0.762**
Informative Feature - o	0.656	0.403	0.491	0.717
Informative Baseline - e	0.639	0.624	0.625	0.746
Informative Baseline - r	0.621	0.589	0.601	0.734
Informative Baseline - k	0.639	0.608	0.620	0.745
Informative Baseline - t	0.633	0.623	0.625	0.745

TABLE VI: Performance of different features. *o*: Odds ratio features, *e*: Emotion features, *r*: Curated n-gram pattern features, *t*: Time features, *k*: keyword-based TF-IDF features

it evokes. Finally, we notice from our evaluations that by combining all of our proposed features we get the best performance.

We next conducted error analysis on false positives and false negatives obtained with the best performing informative baseline and found the following common errors:

Lack of Co-reference Resolution: Users often mention the medication and then indicate they have stopped taking it in the sentences that follow. Consider the following snippet: *Hey guys I took Cymbalta for about three weeks and other than feeling good for the first week I felt like I didn't feel any better. Anyway I stopped it without tapering about 4 days ago and I have been itching all over since then.* We see that the pronoun 'it' refers to 'Cymbalta.' However, this has to be solved automatically.

Lack of Dependency Relations: Dependency parsers help us in understanding constraints that may not be straightforward to encode. As an example, the verb 'taken' refers to all the medications in the following sentence: *I have taken Lexapro, Prozac, Zoloft and now Celexa.* Similarly, we have a sentence *Needless to say, I stopped taking the Cymbalta immediately and have now been off it 7 days.* where the word 'stop' is a clausal complement of the verb 'taking.' These are not captured by the informative baseline features, but may be captured by more advanced feature sets.

Unaccounted Frequent Text Patterns: There were various patterns that seemed to appear frequently that were not captured by our basic informative feature sets. As an example, *I was on zoloft and switched over to paxil for about a year.* Here 'was on Zoloft' is a frequently occurring pattern that we did explicitly capture. However, the complex features we introduce in the next section do capture this type of phrases.

C. Complex Features: Definitions and Examples

Based on the error analysis we have conducted, we propose here three new complex features and use them to further improve the performance of our classifier as well as our understanding of this challenging task.

Dependency Relation Features: Based on the error analysis we have conducted in the previous section, we saw that dependencies between words have a big impact on the classification task. Based on this intuition, we employed the Stanford dependency parser [23] which gives the grammatical relationship between a head word and its modifier word. In particular, this allows us to capture the relationship between a medication word and any other words that may be describing it. Capturing multi-word phrases becomes possible as well, for instance, 'Stopped taking' in our previous example which is given a clausal complement label.

Frequent Sequential Pattern Feature: We further noticed that sentences where patients described the times when they stopped taking medications, have similar structures. In order to capture this, we utilized a frequent sequential pattern mining algorithm [24]. The benefit of capturing frequent sequential patterns as opposed to n-grams is that the algorithm can capture words that are farther apart from each other. As an example, *He put me on effexor.* cannot be captured by a simple bigram (put on) since there is a pronoun that appears in between the two words, whereas this can be captured by frequent sequential patterns (as *put on medication*). In running frequent sequential patterns, we tokenized all occurrences of medication as either the current medication of interest (as 'MEDICATION_TOK'), or other medications that are present in the post, but not something that we are interested in classifying whether the patient has stopped taking or not (as 'OTHER_MEDICATION_TOK'), and set the weight as the log of the number of occurrences the pattern had in our corpus.

Co-reference Resolution Features: In many sentences, people would mention a medication and indicate if they stopped taking it the sentence after. Despite the usage of an existing off-the-shelf co-reference resolution toolkit [25], it quickly became apparent that the out-of-the-box features were not accurate in deciphering the coreferences. Thus, a heuristics method was used to resolve the problem. Whenever we saw words that may be indicative of medications (such as 'medication,' 'med,' and 'anti-depressant' as well as 'it') we replaced this with the last medication that we have seen within a two sentence window. We then could apply all the other features to this recovered co-reference as well.

Summaries of the features described in this section are shown in Table II.

D. Complex Features: Experiments and Analysis

We next experimented with complex features as well. We set the best performing combination of features from informative baseline features as the starting point of this experiment. For all possible combinations we have ran the experiment. Our results are shown on Table VII.

Method	Precision	Recall	F-1	Accuracy
Informative Baseline	0.673	0.640	0.650	0.762
Informative Baseline + d	0.702	0.616	0.652	0.775
Informative Baseline + f	0.768	0.633	0.693	0.809
Informative Baseline + c	0.661	0.634	0.643	0.759
Informative Baseline + d + f	0.756	0.627	0.684	0.802
Informative Baseline + c + f	0.771	0.638	0.695	0.810
Informative Baseline + c + d	0.700	0.606	0.646	0.772
Informative Baseline + d + f + c	**0.786**	**0.636**	**0.701**	**0.817**

TABLE VII: Performance of complex features. *d*: dependency parser features, *f*: frequent sequential pattern features, *c*: co-reference features

We can see the complex features we proposed improved both precision and F-1 score. We have further noticed that frequent sequential pattern features had significant boost in improving the performance. Furthermore, the new features we proposed had improved primarily on precision with small reduction on recall. It is also interesting to note that adding co-reference features, by itself, does not improve performance compared to the informative baseline. However, we see an improvement when the co-reference feature is combined with other features.

There were also cases where the complex features did not perform very well. One common error was when the forum post had sequences of events. As an example, *I got on **Paxil** 2 years ago and switched to Wellbutrin, then I was pregnant and switched to Prozac. I got back on medication, **Paxil**,two months ago.* requires a temporal representation of taking/stop taking the medication, for a given user. A simple 'stop taking' label is difficult because the patient did stop taking the medication at one point, though now the person is taking the medication. While there is no previous research on exactly the problem we tackled here, there is some body of research on generating timelines [26], [27] which we will be able to leverage in future research to mitigate this problem.

Another common error was when the act of stop taking the medication was not explicitly mentioned. From the sentence, *He's been on **Zoloft** but he said that he walked around in a daze all the time and could barely function at work (his job requires that he is alert and ready to respond to problems constantly.) Now he's trying a different med (he takes it sporadically - can't remember the name), and it also makes him tired and like he's in a fog.* We see that the target person has now stopped taking the medication. However, this is inferred from what is written, and one way to classify this is by knowing that by taking a different medication, the person is no longer taking the other medication - i.e., deeper semantic inference. However, there are linguistic devices that will tell us if a person takes multiple pills at the same time, or has switched medication. We will explore such contexts in future work.

V. Discussion

In this section, we analyzed the coefficients from our trained SVM. We have trained SVM using its primal form so that interpreting weights would be easier than if we had used the dual form. The differences in performance were negligible.

Curated n-gram pattern features were a strong indicator, where we found that a lot of patterns which followed *[past tense verb] on medication*, and *[present tense verb] off medication* were associated with 'Stop taking' patterns. Some other patterns that do not involve the prepositions/particles 'on' or 'off' were: 'tried with medication,' 'started with medication,' 'told that medication,' and 'did like medication.' Furthermore, time related features did have a small impact as well, though not as significant as that of the rules. The top 10 important curated n-gram pattern features we discovered are shown in Table VIII.

As for the Emotion features, 'Stop taking' label had strong inverse correlation among aptitude (0.011), pleasantness (0.00684), attention (0.00526) and sensitivity (0.00291) emotional aspects according to SenticNet. These correspond to disgust, sadness, surprise and fear when patients had stopped taking medications. We note that because of the emotion model we used, there are four different dimensions, and hence, four different emotions for each label. These seem to indicate that when patients stop taking medications, there are, indeed, correlations related to emotions, in particular, that of disgust and sadness.

Frequent sequential patterns dominated most of the highest scoring weights. The algorithm is able to capture many of the frequent patterns that occur. However, they may not make immediate semantic sense for the human analyst. In 'Stop taking' labels, past tense verbs ('tried', 'been', and 'was'), and some prepositions/particles ('off') were commonly extracted as part of the patterns, whereas for Others label, present tense verbs ('taking', 'am') and other type of prepositions/particles ('on') were common. For both labels, the current medication of interest (such as Zoloft, Paxil or Cymbalta), annotated as MEDICATION_TOK were usually present in frequent patterns. These findings are consistent with what we have found in curated n-gram pattern features - for example, usefulness of prepositions/particles such as 'off' and 'on,' and phrases that start with 'tried' or 'start.' There were instances where 'MEDICATION_TOK' were not present, but the algorithm still thought it was a frequent pattern. For instance the pattern 'i been off' was considered to be an important feature for the 'Stop taking' label. While this may be a source of ambiguity (in the given frequent pattern, 'i been off,' there is no indication of which medication the author has been on), since a given sentence may be captured by more than one frequent sequential pattern, the fact that some frequent features did not have 'MEDICATION_TOK' is not a problem.

Another interesting observation is that the algorithm was able to find medications that are not the current medication of interest (annotated as 'OTHER_MEDICATION_TOK') to be useful as well, in particular for the Stop_ taking label. For example, we found a very frequent pattern, *I MEDICATION_TOK OTHER_MEDICATION_TOK* (used in the context of, ***I** tried **Zoloft**, **Paxil**, and Cymbalta* if the medication of interest was 'Zoloft.') to be indicative of stop taking medication. This indicates that often times patients mention other medications when they indicate that they have stopped taking a given medication, often times as an enumeration, or that they are now currently taking other medication. The top 10 most important features used by the classifier on both target classes are shown in Table IX.

We used a dependency parser to capture the relationship between words in a sentence with the hope that it would capture relevant relationships between words. Again, our top 10 most important features are shown in Table X. We discovered some verbs that describe medications, such as tapering/VBG - prep - MEDICATION_TOK/NN, and phrases that indicate (but do not necessarily explic-

Stop Taking	Others
was/VBD on/IN MEDICATION TEMPORAL_WORD	have/VBP on/IN MEDICATION TEMPORAL_WORD
have/VBP off/RP MEDICATION TEMPORAL_WORD	've/VBP on/IN MEDICATION TEMPORAL_WORD
was/VBD on/IN MEDICATION	am/VBP on/IN MEDICATION
had/VBD on/IN MEDICATION TEMPORAL_WORD	put/VBD on/IN MEDICATION MEDICATION
have/VBP off/IN MEDICATION TEMPORAL_WORD	have/VBP on/IN MEDICATION
've/VBP off/IN MEDICATION	put/VBD on/IN MEDICATION
got/VBD on/IN MEDICATION MEDICATION	'm/VBP on/IN MEDICATION MEDICATION
am/VBP off/RP MEDICATION TEMPORAL_WORD	started/VBD on/IN MEDICATION TEMPORAL_WORD
've/VBP off/RP MEDICATION TEMPORAL_WORD	'm/VBP on/IN MEDICATION
was/VBD like/IN MEDICATION	got/VBD on/IN MEDICATION TEMPORAL_WORD

TABLE VIII: Top 10 important curated n-gram pattern features found by the classifier

Stop Taking	Others
i MEDICATION_TOK OTHER_MEDICATION_TOK	i taking MEDICATION_TOK
i on OTHER_MEDICATION_TOK	i am MEDICATION_TOK
i MEDICATION_TOK again	been MEDICATION_TOK for
i MEDICATION_TOK off	i of MEDICATION_TOK
i off MEDICATION_TOK	i like MEDICATION_TOK
i and OTHER_MEDICATION_TOK	i just MEDICATION_TOK
off MEDICATION_TOK a	i MEDICATION_TOK mg
i been off	my me MEDICATION_TOK
i on MEDICATION_TOK but	for and MEDICATION_TOK
tried MEDICATION_TOK and	MEDICATION_TOK and if

TABLE IX: 10 most important frequent sequential features the classifier used

itly mention) stop taking the medications, such as withdrawal/NN - nn - cymbalta/NNP to be associated with the Stop_taking label. On the other hand, dependency relations such as taking/VBG - dobj - MEDICATION_TOK/NNP or prescribed/VBN - dobj - MEDICATION_TOK/NN were associated with 'Others' label. Some of the patterns we found here were also present in frequent sequential patterns. For instance, the algorithm found MEDICATION_TOK/NNP - conj - OTHER_MEDICATION_TOK/NNP to be a useful pattern which is what we have also found via frequent sequential patterns. However, as was shown in our evaluation results, there were enough patterns that the frequent sequential patterns did not capture but were captured by the dependency parser.

We analyzed keyword-based features, and focused our attention on medications that were correlated with 'Stop taking' label or 'Others' label. Other than the medications that we focused on (Paxil, Zoloft and Cymbalta), we found that Lexapro, Prozac, Wellbutrin, Effexor, Celexa, Lamictal, Remeron, Abilify, and Pristiq were associated with 'Stop taking' labels as well. This seems to indicate again that patients often mention multiple medications when they indicate that they have stopped taking a given medication. On the other hand, the medications Seroquel and Buspar had weak negative correlations with the 'Stop taking' label. Seroquel is primarily used as an antipsychotic, and often used in conjunction with other medications to treat depression, and Buspar is often used to treat anxiety. Neither of the medications were primarily used as anti-depressants, which may explain why they had weak negative correlations to the 'Stop taking' label. However, we leave these issues for a future in-depth analysis of the complex problem of stop taking medications in social media forums.

VI. Conclusion

In this paper we looked at different features of various compelxity to learn more about the problem domain. We started out by taking odds ratio of the words that appear in our corpus (i.e., 'baseline'). We then proposed adding more features motivated by existing approaches that seemed to perform well in similar problem domains and tasks (i.e., 'informed baseline'). After conducting an error analysis, we went on to propose complex features meant to capture some of the errors that the existing features were unable to capture - and these significangly improved the performance. As future work, we have several directions. First, we would like to investigate on how to encode the current state of the medication. Some of the common errors that the classifier made were in cases where there were multiple instances of the medication of interest. By constructing these into a timeline, we would be able to capture what the current state of the medication prescription may be. Furthermore, both our keyword-based features and curated n-gram pattern features, while effective at generating relevant phrases, would benefit from a more automated method to increase coverage. We leave this as future work to further increase recall. Another direction, which is an extension of the previous direction in encoding the current status of the medication is by adding a new label, 'currently taking medication' as well as constraints to help us classify whether the person has stopped taking the medication or not. We noticed that a person is not likely to take more than one or two medications at a same time, especially if it pertains to the same class of medication (although this is possible). By studying this constraint, we would be able to better classify whether the person has stopped taking the medication or not. Finally, we would like to investigate the

Stop Taking	Others
longer/RB - neg - no/RB	taking/VBG - dobj - MEDICATION_TOK/NNP
taken/VBN - nsubj - i/PRP	day/NN - det - a/DT
stopped/VBD - xcomp - working/VBG	take/VBP - nsubj - i/PRP
taking/VBG - aux - was/VBD	wondering/VBG - nsubj - i/PRP
had/VBD - neg - never/RB	taking/VBG - dobj - cymbalta/NNP
stopped/VBD - nsubj - i/PRP	have/VBP - nsubj - i/PRP
years/NNS - amod - few/JJ	started/VBD - advmod - just/RB
was/VBD - nsubj - i - P/RP	ago/RB - npadvmod - months/NNS
tried/VBN - nsubj - i/PRP	tell/VB - aux - can/MD
had/VBD - nsubj - i/PRP	'm/VBP - nsubj - i/PRP

TABLE X: Top 10 most important dependency features found by the classifier

reasons for which a patient has stopped taking the medication. This would provide useful information to researchers working on healthcare problems or medical doctors on the type of medication the patient should be taking.

REFERENCES

[1] W. W. Fleischhacker, M. A. Oehl, and M. Hummer, "Factors influencing compliance in schizophrenia patients," *J Clin Psychiatry*, vol. 64 Suppl 16, pp. 10–13, 2003.

[2] H. Rittmannsberger, T. Pachinger, P. Keppelmuller, and J. Wancata, "Medication adherence among psychotic patients before admission to inpatient treatment," *Psychiatr Serv*, vol. 55, no. 2, pp. 174–179, Feb 2004.

[3] M. R. DiMatteo, "Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research," *Med Care*, vol. 42, no. 3, pp. 200–209, Mar 2004.

[4] D. Roe, H. Goldblatt, V. Baloush-Klienman, M. Swarbrick, and L. Davidson, "Why and how people decide to stop taking prescribed psychiatric medication: exploring the subjective process of choice," *Psychiatr Rehabil J*, vol. 33, no. 1, pp. 38–46, 2009.

[5] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance," *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct 2015.

[6] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic," *PLoS ONE*, vol. 6, no. 5, p. e19467, 2011.

[7] S. Wang, Y. Li, D. Ferguson, and C. Zhai, "Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB '14. New York, NY, USA: ACM, 2014, pp. 321–330.

[8] O. Uzuner, I. Goldstein, Y. Luo, and I. Kohane, "Identifying patient smoking status from medical discharge records," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 14–24, 2008.

[9] P. J. McCormick, N. Elhadad, and P. D. Stetson, "Use of semantic features to classify patient smoking status," *AMIA Annu Symp Proc*, pp. 450–454, 2008.

[10] D. L. MacLean, S. Gupta, A. Lembke, C. D. Manning, and J. Heer, "Forum77: An analysis of an online health forum dedicated to addiction recovery," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, 2015, pp. 1511–1526.

[11] R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, and N. H. Shah, "Text mining for adverse drug events: the promise, challenges, and state of the art," *Drug Saf*, vol. 37, no. 10, pp. 777–790, Oct 2014.

[12] B. W. Chee, R. Berlin, and B. Schatz, "Predicting adverse drug events from personal health messages," *AMIA Annu Symp Proc*, vol. 2011, pp. 217–226, 2011.

[13] M. J. Paul, M. S. Chisolm, M. W. Johnson, R. G. Vandrey, and M. Dredze, "Assessing the Validity of Online Drug Forums as a Source for Estimating Demographic and Temporal Trends in Drug Use," *J Addict Med*, Jul 2016.

[14] J. H. D. Cho, V. Q. Z. Liao, Y. Jiang, and B. R. Schatz, "Aggregating personal health messages for scalable comparative effectiveness research," in *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM-BCB 2013, Washington, DC, USA, September 22-25, 2013*, 2013, p. 907.

[15] R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, and E. Horvitz, "Toward enhanced pharmacovigilance using patient-generated data on the internet," *Clin. Pharmacol. Ther.*, vol. 96, no. 2, pp. 239–246, Aug 2014.

[16] G. N. Noren, "Pharmacovigilance for a revolving world: prospects of patient-generated data on the internet," *Drug Saf*, vol. 37, no. 10, pp. 761–764, Oct 2014.

[17] L. Andrade, J. J. Caraveo-Anduaga, P. Berglund, R. V. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R. C. Kessler, N. Kawakami, C. Kilic, D. Offord, T. B. Ustun, and H. U. Wittchen, "The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys," *Int J Methods Psychiatr Res*, vol. 12, no. 1, pp. 3–21, 2003.

[18] E. Aramaki, T. Imai, K. Miyo, and K. Ohe, "Patient Status Classification by Using Rule based Sentence Extraction and BM25 kNN-based Classifier," *Processing for Clinical Data*, 2006.

[19] A. M. Cohen, "Five-way smoking status classification using text hot-spot identification and error-correcting output codes," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 32–35, 2008.

[20] R. Wicentowski and M. R. Sydes, "Using implicit information to identify smoking status in smoke-blind medical discharge summaries," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 29–31, 2008.

[21] C. Clark, K. Good, L. Jezierny, M. Macpherson, B. Wilson, and U. Chajewska, "Identifying smokers with a medical extraction system," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 36–39, 2008.

[22] E. Cambria, D. Olsher, and D. Rajagopal, "Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 2014, pp. 1515–1521.

[23] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 740–750.

[24] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining sequential patterns by prefix-projected growth," in *Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany*, 2001, pp. 215–224.

[25] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 294–303.

[26] J. Li and C. Cardie, "Timeline generation: Tracking individuals on twitter," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 643–652.

[27] Z. Chen and H. Ji, "Language specific issue and feature exploration in chinese event extraction," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 209–212.